Radiology

Development and Validation of a Deep Learning System for Staging Liver Fibrosis by Using Contrast Agent–enhanced CT Images in the Liver

*Kyu Jin Choi, BS** • *Jong Keon Jang, MD** • *Seung Soo Lee, MD* • *Yu Sub Sung, PhD* • *Woo Hyun Shim, PhD* • *Ho Sung Kim, MD* • *Jessica Yun, BS* • *Jin-Young Choi, MD* • *Yedaun Lee, MD* • *Bo-Kyeong Kang, MD* • *Jin Hee Kim, MD* • *So Yeon Kim, MD* • *Eun Sil Yu, MD*

From the Department of Computer Science, Hanyang University, Seoul, Republic of Korea (K.J.C.); Department of Radiology and Research Institute of Radiology (J.K.J., S.S.L., Y.S.S., W.H.S., H.S.K., J.Y., J.H.K., S.Y.K.) and Department of Diagnostic Pathology (E.S.Y.), Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, South Korea; Department of Radiology, Severance Hospital, Yonsei University College of Medicine, Seoul, Korea (J.Y.C.); Department of Radiology, Haeundae Paik Hospital, Inje University College of Medicine, Busan, Korea (Y.L.); and Department of Radiology, Hanyang University Medical Center, Hanyang University School of Medicine, Seoul, Korea (B.K.K.). Received April 3, 2018; revision requested May 15; revision received July 8; accepted July 17. **Address correspondence to** S.S.L. (e-mail: *seungsoolee@amc.seoul.kr*).

Study supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, Information and Communications Technology (ICT) and Future Planning (NRF-2017R1A2B4003114), and the Bio and Medical Technology Development Program of the NRF funded by the Ministry of Science and ICT (NRF-2016M3A9A7918706).

*K.J.C. and J.K.J. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the editorial by Smith in this issue.

Radiology 2018; 289:688–697 • https://doi.org/10.1148/radiol.2018180763 • Content codes: IN GI CT

Purpose: To develop and validate a deep learning system (DLS) for staging liver fibrosis by using CT images in the liver.

Materials and Methods: DLS for CT-based staging of liver fibrosis was created by using a development data set that included portal venous phase CT images in 7461 patients with pathologically confirmed liver fibrosis. The diagnostic performance of the DLS was evaluated in separate test data sets for 891 patients. The influence of patient characteristics and CT techniques on the staging accuracy of the DLS was evaluated by logistic regression analysis. In a subset of 421 patients, the diagnostic performance of the DLS was compared with that of the radiologist's assessment, aminotransferase-to-platelet ratio index (APRI), and fibrosis-4 index by using the area under the receiver operating characteristic curve (AUROC) and Obuchowski index.

Results: In the test data sets, the DLS had a staging accuracy of 79.4% (707 of 891) and an AUROC of 0.96, 0.97, and 0.95 for diagnosing significant fibrosis (F2–4), advanced fibrosis (F3–4), and cirrhosis (F4), respectively. At multivariable analysis, only pathologic fibrosis stage significantly affected the staging accuracy of the DLS ($P = .016$ and .013 for F1 and F2, respectively, compared with F4), whereas etiology of liver disease and CT technique did not. The DLS (Obuchowski index, 0.94) outperformed the radiologist's interpretation, APRI, and fibrosis-4 index (Obuchowski index range, 0.71–0.81; $P < .001$) for staging liver fibrosis.

Conclusion: The deep learning system allows for accurate staging of liver fibrosis by using CT images.

Online supplemental material is available for this article.

Fibrosis of the liver is an important cause of morbidity and mortality in patients with chronic liver disease, and patients with advanced fibrosis are likely to develop complications. Although liver biopsy is the current reference standard for assessing liver fibrosis, it is associated with a risk of procedure-related complications and some limitations, such as sampling error and interobserver variability (1,2).

A number of noninvasive methods have been investigated as alternatives for liver biopsy (3–9), of which measurement of liver stiffness by using US or MR elastography has been the most successfully implemented in clinical practice (6,7). Liver fibrosis is accompanied by macroscopic and microscopic morphologic changes in the liver that are reflected on US, CT, or MR images as a blunted liver margin, surface nodularity, and coarse texture (4,10–13). Visual assessment of these imaging features is used to detect liver cirrhosis in clinical practice (12–14). However, visual assessment has low sensitivity, especially for detecting fibrosis in the early stages, and also exhibits interobserver variability because of the subjective nature of visual analysis of images (12,15–18). Recently, quantitative methods for analysis of liver surface nodularity at imaging examination showed promising results (10,12,19) and allowed for accurate staging of liver fibrosis within an acceptable operating time.

Deep learning on the basis of a convolutional neural network recently attracted attention as a method of image recognition and interpretation in medicine (20–24). By using a supervised learning process that included a large data set of labeled images, a deep learning system (DLS) can be trained for use in clinical decision making. We hypothesized that it would be possible to develop a DLS for automated staging of liver fibrosis. Because of

Abbreviations

APRI = aminotransferase-to-platelet ratio index, AUROC = area under the receiver operating characteristic curve, CI = confidence interval, DLS = deep learning system, METAVIR = Meta-analysis of Histological Data in Viral Hepatitis

Summary

A deep learning system developed by using a large development data set allowed for accurate assessment of liver fibrosis with portal venous phase CT images in separate test data sets.

Implication for Patient Care

A CT-based deep learning system may be an accurate and widely applicable clinical tool for assessing liver fibrosis by using routine portal venous phase CT images of the liver.

the wide availability of CT and its frequent use in evaluation of patients with various types of chronic liver disease, automated staging of liver fibrosis based on CT images would have wide applicability. Furthermore, a DLS on the basis of automated liver segmentation might avoid interreader variability in assessment of liver fibrosis compared with other quantitative methods that involve selection of regions of interest by radiologists (10,12,19).

The purpose of this study was to develop and validate a DLS for staging liver fibrosis by using portal venous phase CT images.

Materials and Methods

The study was approved by the institutional review boards of the four participating institutions. The requirement for informed consent was waived because of the retrospective nature of the data analysis.

Development Data Set

We used a data set of contrast agent–enhanced portal venous phase liver CT images in 7461 patients with pathologic examination–confirmed liver fibrosis to develop a DLS for staging of liver fibrosis. The data set was derived from 12 535 patients who underwent liver resection, liver transplant, or liver biopsy and contrast-enhanced portal venous phase liver CT within 3 months of pathologic examination of the liver at Asan Medical Center from 2007 to 2016. The inclusion criteria were as follows: age, 18 years or older; availability of pathologic reports including stage of hepatic fibrosis; no previous liver surgery; no anticancer treatment within 6 months of pathologic examination of the liver; and no hepatic tumor larger than 5 cm in diameter. Of the 7889 eligible patients who underwent US-guided percutaneous liver parenchymal biopsy, 428 patients were assigned to test data set 1, and the development data set was composed of data for the remaining 7461 patients (Fig 1). The characteristics of the development data set are described in Table 1.

Test Data Sets for Clinical Validation

Three independent test data sets that included data from 891 patients with pathologic examination–confirmed liver fibrosis were used for clinical validation of the DLS. Test data set 1 was derived from the same eligible patients as in the development data set. To represent a clinically relevant condition in which assessment of liver fibrosis is required, patients who underwent US-guided percutaneous liver parenchymal biopsy for evaluation of abnormal liver function test results or who were suspected of having or who had a diagnosis of chronic liver disease were selected from the 7889 patients who met the inclusion criteria for the development data set. Of the 428 patients selected, seven patients whose biopsy results indicated liver involvement of lymphoma (n = 5) or amyloidosis (n = 2) were excluded. Test data set 1 was composed of data for the remaining 421 patients.

For external validation of the DLS, test data sets 2 and 3 were obtained at a variety of clinical situations by using variable CT techniques and from multiple outside institutions. Test data set 2 (n = 298) was obtained from the patients who underwent liver resection, liver transplantation, or liver biopsy at Asan Medical Center from 2007 to 2016, who had liver CT scan performed at outside institutions within 3 months of pathologic examination of the liver, who met the same inclusion criteria as the development data set, and who were not included in the development data set or test data set 1. Test data set 3 (n = 172) consisted of data obtained from three tertiary referral hospitals (Inje University Paik Hospital, Hanyang University Hospital, and Yonsei University Severance Hospital). The inclusion criteria for test data set 3 were as follows: age, 18 years or older; pathologic examination–confirmed liver fibrosis; no previous liver surgery; and CT data acquired within 5 months of pathologic examination of the liver. The data from Inje University Paik Hospital (n = 40) and Hanyang University Hospital (n = 40) were derived from the patients who underwent US-guided percutaneous liver biopsy for evaluation of abnormal liver function test results or for suspected or known chronic liver disease from 2014 to 2017. The data (n = 92) from Yonsei University Severance Hospital were from patients who underwent resection of hepatic tumors from June 2010 to December 2011; these patients formed part of the study population in a previously published study (11) that investigated histogram analysis of MR images for staging liver fibrosis. Our study used CT data of these patients to validate the DLS. Flow diagrams for test data sets are in Figure 1. The detailed characteristics of the test data set are described in Table 1.

CT Examination

The CT data in this study were collected over a long period and from multiple institutions, and various CT techniques were used (Table E1 [online]). Most CT images were obtained by using 16 (or higher) multi–detector row CT systems (6944 of 7461 examinations [93.1%] in the development data set and 750 of 891 examinations [84.2%] in the test data sets). Portal venous phase imaging was performed at 70–80 seconds after intravenous administration of a contrast agent. Most CT images were acquired at 120 kVp (n = 6445 [86.4%] in the development data set, n = 791 [88.8%] in the test data sets) and reconstructed with a 5-mm

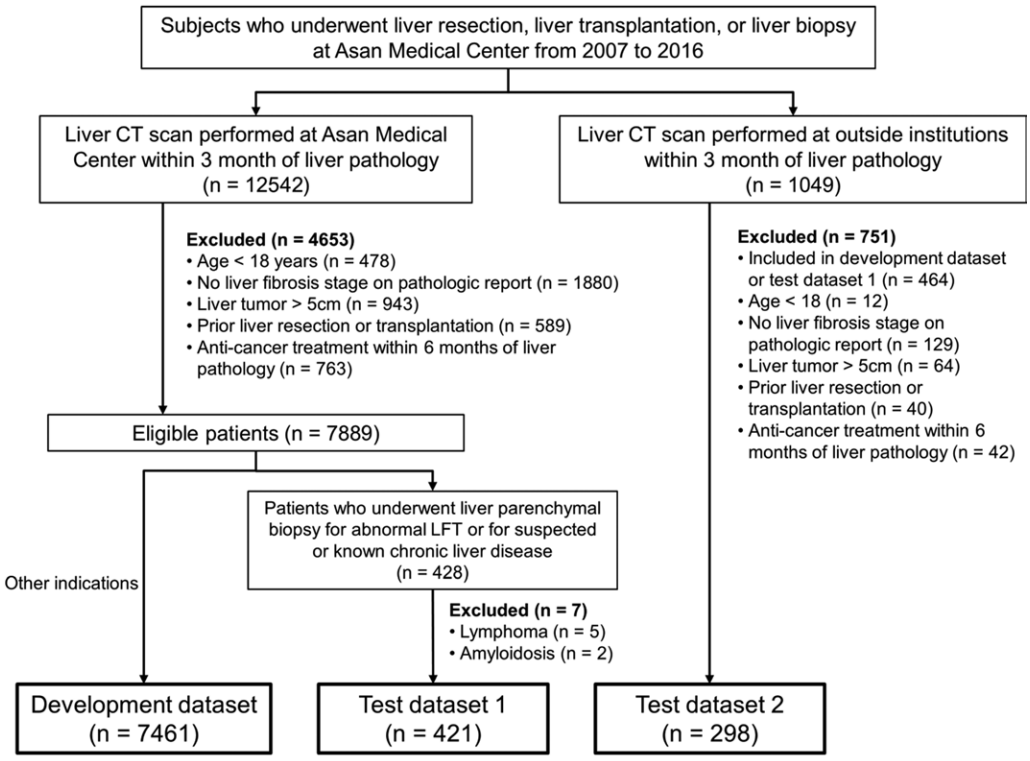

Figure 1: Flow diagrams for the development and test data sets. **(a)** The development data set and test data set 1 included the CT and liver pathology data obtained at Asan Medical Center. The test data set 2 consisted of CT data obtained at the various institutions and pathologic data obtained at Asan Medical Center. **(b)** Test data set 3 consisted of CT and pathology data from the three institutions. LFT = liver function test.

section thickness (n = 5309 [71.2%] in the development data set, n = 655 [73.5%] in the test data sets).

Reference Standard for Liver Fibrosis

Liver pathologic examination served as the reference standard for liver fibrosis. Pathologic assessment of liver fibrosis was performed by using hematoxylin-eosin and Masson trichrome staining according to the guideline for histologic grading and staging of chronic hepatitis proposed by the Korean Study Group for the Pathology of Digestive Diseases (25), which uses the same liver fibrosis staging system as the Meta-analysis of Histological Data in Viral Hepatitis (METAVIR) fibrosis staging system (26) as follows: F0, no fibrosis; F1, portal fibrosis; F2, periportal fibrosis; F3, septal fibrosis; and F4, cirrhosis.

DLS Development

The data set used to develop the DLS was imbalanced: the amount of data for pathologic liver fibrosis stages was smaller for F1, F2, and F3 than for F0 and F4. This imbalance could have resulted in poor classification accuracy by the DLS for the minority classes (ie, there could be a tendency to classify a test sample in one of the majority classes). To overcome this problem, we augmented the image data for the minority classes (F1, F2, and F3) by rotating the original CT images between −15° and 15° and/or by adding Gaussian noise (at a σ level ranging from 0.9 to 1.1) on a random basis so that the amount of imaging data used for development of the DLS was balanced across all fibrosis stages. The results of the preliminary experiment to evaluate the effect of augmentation of the image data on the performance of the DLS are shown in Appendix E1 (online).

Our DLS for CT-based liver fibrosis staging consists of two distinct algorithms (ie, one for liver segmentation and the other for staging of liver fibrosis), both of which are on the basis of a convolutional neural network.

The algorithm for liver segmentation was trained by using liver outlines drawn with software (ImageJ; National Institutes of Health, Bethesda, Md) by an experienced radiologist (S.S.L., with 15 years of experience in abdominal radiology) on portal venous phase CT images for 50 patients randomly selected from the development data set. The algorithm for liver segmentation

Table 1: Characteristics of Development and Test Data Sets

Characteristic	Development Data Set	Test Data Set			
		Total	Data Set 1	Data Set 2	Data Set 3
No. of Patients	7461	891	421	298	172
Age (y)*	44.2 ± 14.7 (18–83)	51.5 ± 13.3 (18–86)	47.9 ± 13.5 (18–81)	55.8 ± 10.3 (18–86)	52.6 ± 14.7 (18–81)
No. of male participants	5385 (72.2)	511 (57.4)	200 (47.5)	222 (74.5)	89 (51.7)
Pathologic liver fibrosis stage					
F0	3357 (45.0)	118 (13.2)	64 (15.2)	17 (5.7)	37 (21.5)
F1	113 (1.5)	109 (12.2)	54 (12.8)	14 (4.7)	41 (23.8)
F2	284 (3.8)	161 (18.1)	108 (25.6)	33 (11.1)	20 (11.6)
F3	460 (6.2)	173 (19.4)	86 (20.4)	65 (21.8)	22 (12.8)
F4	3247 (43.5)	330 (37.0)	109 (25.9)	169 (56.7)	52 (30.2)
Etiologic cause of liver disease					
HBV	2995 (40.1)	399 (44.8)	90 (21.4)	249 (83.6)	60 (34.9)
HCV	374 (5.0)	102 (11.4)	76 (18.1)	17 (5.7)	9 (5.2)
Alcohol	494 (6.6)	43 (4.8)	24 (5.7)	11 (3.7)	8 (4.7)
Autoimmune†	38 (0.5)	152 (17.1)	119 (28.3)	4 (1.3)	29 (16.9)
Other‡	59 (0.8)	122 (13.7)	84 (20.0)	2 (0.7)	36 (20.9)
Unknown	126 (1.7)	58 (6.5)	28 (6.7)	7 (2.3)	23 (13.4)
No§	3375 (45.2)	15 (1.7)	0 (0)	8 (2.7)	7 (4.1)
Pathology specimen					
US-guided biopsy	3386 (45.4)	509 (57.1)	421 (100)	8 (2.7)	80 (46.5)
US-guided liver parenchymal biopsy	3328 (44.6)	504 (56.6)	421 (100)	3 (1.0)	80 (46.5)
Liver resection	2013 (27.0)	379 (42.5)	0 (0)	287 (96.3)	92 (53.5)
Transplant	2062 (27.6)	3 (0.3)	0 (0)	3 (1.0)	0 (0)
Liver neoplasm					
Absent	4645 (62.3)	506 (56.8)	421 (100)	7 (2.3)	78 (45.3)
Present	2816 (37.7)	385 (43.2)	0 (0)	291 (97.7)	94 (54.7)
HCC	2510 (33.6)	334 (37.5)	0 (0)	269 (90.3)	65 (37.8)
Other malignancy	167 (2.2)	42 (4.7)	0 (0)	17 (5.7)	25 (14.5)
Benign tumor	139 (1.9)	9 (1.0)	0 (0)	5 (1.7)	4 (2.3)
Interval between CT and pathologic analysis (d)*	10.8 ± 15.2 (0–90)	11.5 ± 20.6 (0–153)	6.7 ± 20.4 (0–90)	12.1 ± 10.0 (0–63)	22.2 ± 29.0 (0–153)

Note.—Unless otherwise indicated, data are number of participants; data in parentheses are percentages. F0 = no fibrosis, F1 = portal fibrosis, F2 = periportal fibrosis, F3 = septal fibrosis, F4 = cirrhosis, HBV = hepatitis B viral, HCC = hepatocellular carcinoma, HCV = hepatitis C viral.

* Data are mean ± standard deviation; data in parentheses are range.

† Included autoimmune hepatitis, autoimmune cholangitis, primary biliary cirrhosis, and primary sclerosing cholangitis.

‡ Included nonalcoholic fatty liver disease, toxic hepatitis, and Wilson disease.

§ Category included donor candidates for living donor liver transplant or patients with hepatic metastasis from other primary malignancy.

was then developed by using these labeled CT data with five-fold cross-validation. The performance of the liver segmentation algorithm, represented by the Dice similarity index equation [2 · true-positive finding/(2 · true-positive finding + false-negative finding + false-positive finding)], was 0.92. The algorithm for staging liver fibrosis was trained by using the entire development data set (composed of 7461 CT examinations). The CT data were first processed by using the previously mentioned liver segmentation algorithm, and the segmented liver data were then entered into the algorithm for staging liver fibrosis. The output data were compared with the pathologic stage of fibrosis and the error was back-propagated (a method to train a deep learning algorithm) to optimize the architecture and parameters of the convolutional neural network. The performance of the DLS in the development data set was evaluated by five-fold cross-validation (by using a ratio of 4:1 for the training and validation sets). A schematic diagram of the DLS is shown in Figure 2. The details of the DLS are described in Appendix E1 (online).

DLS Evaluation

Clinical validation of the DLS was performed by using the test data sets. The trained DLS extracts the liver region for a given patient from the input CT data, processes the data for the liver, and returns the probability of each stage of fibrosis. The fibrosis stage for which the probability value was highest was assigned as the fibrosis stage determined by the DLS. The staging accuracy of the DLS was calculated as the number of fibrosis stages made by the DLS that were concordant with the pathologic fibrosis stages divided by the number of all tested cases. The accuracy of the DLS in making binary decisions for diagnosing

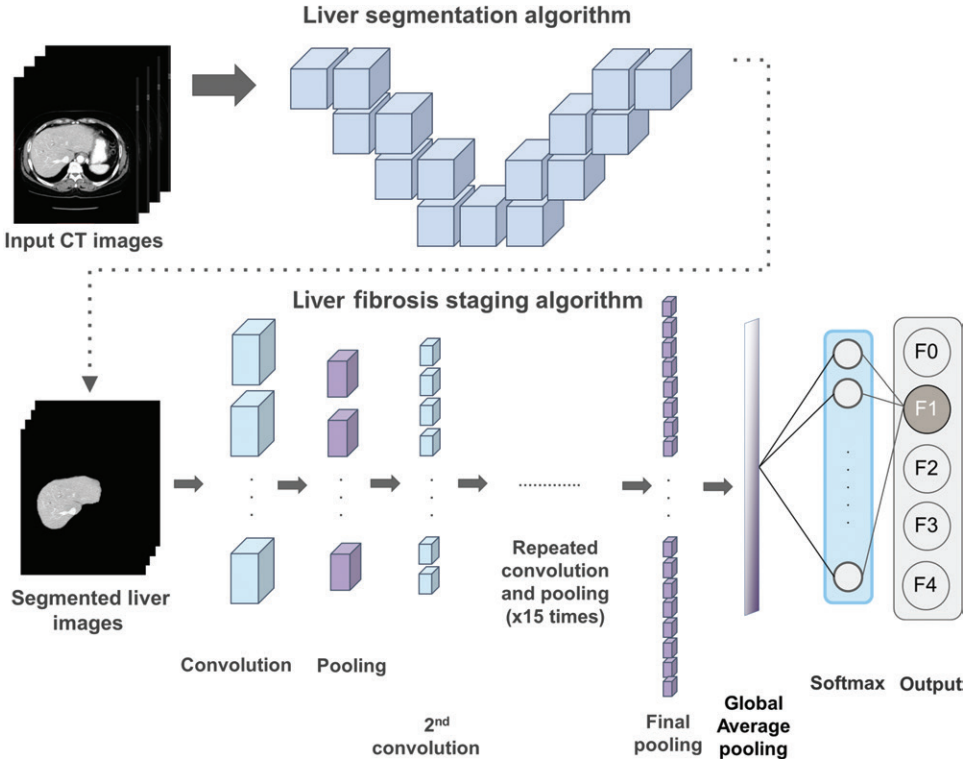

Figure 2: Schematic diagram of a deep learning system for staging of liver fibrosis. The input CT images are first processed with the algorithm for liver segmentation. Segmented liver images are then put into the algorithm for liver fibrosis staging that was constructed by using a three-dimensional convolutional neural network. Repeated convolution and pooling of layers extracts multiple feature maps from the input images by progressively increasing the number of feature maps and reducing their size. The global average pooling layer produces a spatial average of the feature maps that is fed into the Softmax layer to generate the final output (ie, the probability for each fibrosis stage). F0 = no fibrosis, F1 = portal fibrosis, F2 = periportal fibrosis, F3 = septal fibrosis, F4 = cirrhosis.

significant fibrosis (F2–F4), advanced fibrosis (F3–F4), and cirrhosis (F4) was also evaluated by using receiver operating characteristic analysis. The sensitivity, specificity, and accuracy of the DLS for these binary decisions were determined by using the F2, F3, and F4 fibrosis stages predicted by the DLS as the cutoffs for diagnosing significant fibrosis, advanced fibrosis, and cirrhosis, respectively.

Comparison of Diagnostic Performance between the DLS, Readers, and Serum Fibrosis Markers

For test data set 1, three academic abdominal radiologists (S.S.L., S.Y.K., and J.H.K., each with more than 10 years of experience) and one trainee abdominal radiology fellow (J.K.J.) blinded to the pathologic fibrosis staging independently reviewed the anonymized CT images by using a Digital Imaging and Communications in Medicine viewer (Radiant; Medixant, Poznan, Poland). The readers graded the degree of liver fibrosis by using a five-point scale (0, normal; 1, probably normal; 2, chronic liver disease; 3, chronic liver disease or cirrhosis; 4, definite cirrhosis) on the basis of the morphologic features of the liver and the finding of portal hypertension on CT images. General instructions regarding the morphologic characteristics of liver cirrhosis and chronic liver disease (Appendix E2 [online]) were provided before the CT images were interpreted. CT images in 50 patients (including 10 patients for each METAVIR fibrosis stage) that were randomly selected from the development data sets were then given to the readers with the pathologic fibrosis stages for reference. Serum fibrosis markers, including the aminotransferase-to-platelet ratio index (APRI) (27) and fibrosis-4 index (28) were calculated as follows: APRI = aminotransferase level (/unit normal range)/platelet count ($\times 10^9$/L); and fibrosis-4 index = (age [year] $\times$ aminotransferase [U/L])/(platelet count [$\times 10^9$/L] $\times$ (alanine aminotransferase [U/L])$^{1/2}$) by using the results of laboratory tests performed within 7 days $\pm$ 12.6 (standard deviation; range, 1–88 days) of liver biopsy.

Statistical Analysis

The staging accuracy values for the DLS (ie, the proportion of correct fibrosis staging performed by the DLS) in the three test data sets were compared by using Fisher exact test. The influence of patient characteristics, CT data, and pathologic data on the performance of the DLS for staging liver fibrosis in the test data sets was evaluated by using univariate and multivariable logistic regression analyses, with the correctness of the fibrosis stage predicted by the DLS as the dependent variable. Variables with a *P* value less than .1 in univariate analysis were included in the multivariable analysis.

In test data set 1, the interreader agreement of the four radiologists' liver fibrosis grades was assessed by using the intraclass correlation coefficient (two-way random effects model; absolute agreement). The performances of the DLS, radiologists, and serum fibrosis tests in diagnosing significant fibrosis, advanced fibrosis, and cirrhosis were evaluated by using the area under the receiver operating characteristic curve (AUROC). The AUROC values were compared by using the method devised by DeLong et al (29). The diagnostic performance was also evaluated by using the Obuchowski index, which is a multinomial version of the AUROC adapted for ordinal references such as pathologic staging of liver fibrosis (30). The Obuchowski index is a weighted average of the AUROC values obtained for all possible pairs of fibrosis stages (ie, 10 pairs for the five [F0–F4] fibrosis stages) to be differentiated, and it estimates the probability that a test will correctly rank two randomly chosen patients with different stages of fibrosis. The

statistical analyses were performed by using software (SPSS version 21.0, IBM, Armonk, NY; and R studio version 1.1.383, R Foundation for Statistical Computing, Vienna, Austria) with statistical code packages pROC for the ROC analysis and ordROC for the Obuchowski index (31). *P* values less than .05 indicated statistical significance, with no correction for type 1 error used.

Results

The training curve for the DLS is in Figure E1 (online). The cross-validated staging accuracy of DLS was 83.1% (6200 of 7461; 95% confidence interval [CI]: 81.1%, 84.9%) in the development data set. The clinical validation results for the test data sets including 891 patients are summarized as a confusion matrix for the liver fibrosis stage predicted by the DLS compared with the pathologic stage of liver fibrosis (Fig 3). The DLS predicted the same liver fibrosis stage as the pathologic fibrosis stage in 707 of 891 patients, resulting in a staging accuracy of 79.4% (707 of 891; 95% CI: 76.5%, 82.0%). One-hundred and fifty-nine (86.4%) of 184 inaccurate results with the DLS were within one stage of difference from the pathologic fibrosis stage. The DLS was noted to have slightly poorer accuracy for test data set 3 (74.4%; 128 of 172) than for test data sets 1 (80.8%; 340 of 421) and 2 (80.2%; 239 of 298), but the difference was not statistically significant (*P* = .85, .10, and .17 for data set 1 vs 2, data set 1 vs 3, and data set 2 vs 3, respectively). The confusion matrix for each test data set is in Figure E2 (online). For binary decisions including the diagnosis of significant fibrosis, advanced fibrosis, and cirrhosis, the DLS achieved AUROC values of 0.95–0.97, sensitivity of 84.6%–95.5%, specificity of 89.9%–96.6%, and accuracy of 92.1%–95.0% (Table 2).

To evaluate the robustness of the DLS across various clinical settings, we evaluated the influence of the patient characteristics, CT data, and pathologic data on the performance of the DLS for liver fibrosis staging by using all the test data sets (Table 3). At univariable analysis, the performance of DLS for staging liver fibrosis was significantly affected by the pathologic fibrosis stage (*P* = .01 for F1 and *P* < .01 for F2, compared with F4) and the etiologic cause of liver disease (*P* = .04 for other or cryptogenic compared with hepatitis B). Patient demographic characteristics (*P* = .51 for age and .80 for sex), presence of a hepatic tumor (*P* = .68), methods used to acquire the pathologic specimen (*P* = .52), sources of CT (*P* = .32) and pathologic data (*P* = .08), and CT techniques (*P* value range, .19–.28) had no significant influence on the performance of the DLS. Multivariable analysis demonstrated that pathologic stage of liver fibrosis was the only independent factor that influenced the performance of the DLS, which indicated that diagnosis of intermediate fibrosis stages by using the DLS (ie, F1 [adjusted odds ratio, 0.51; *P* = .02]; and F2 [adjusted odds ratio, 0.54; *P* = .01]) was associated with lower accuracy compared with diagnosis of cirrhosis.

A comparison of the performance of the DLS, the radiologists' assessments, and the serum fibrosis tests for assessing liver fibrosis in test data set 1 is in Table 4 and Figure 4. The intraclass correlation coefficient for agreement between the four radiologists' grading results was 0.93 (95% CI: 0.92, 0.94). For

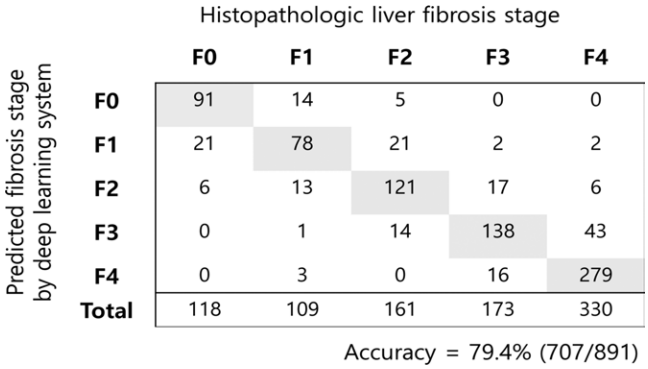

Figure 3: Confusion matrix showing the results for the deep learning system in comparison with pathologic liver fibrosis staging in the complete test data sets. The shaded cells indicate the correct results obtained by the deep learning system.

the diagnosis of significant fibrosis, advanced fibrosis, and cirrhosis, the DLS (AUROC range, 0.95–0.97) outperformed the four radiologists (AUROC range, 0.75–0.88) and the two serum fibrosis tests (AUROC range, 0.65–0.85; *P* < .01 for all comparisons). Similarly, the Obuchowski index value was significantly higher for the DLS (Obuchowski value, 0.94) compared with the values for the radiologists (Obuchowski value range, 0.74–0.81), APRI (Obuchowski value, 0.71), and fibrosis-4 index (Obuchowski value, 0.76; *P* < .01 for all comparisons). The detailed results of pairwise comparisons of the AUROC and Obuchowski index values are summarized in Table E2 (online). Representative cases are in Figure 5.

Discussion

In this study, we sought to develop and evaluate a DLS for automated staging of liver fibrosis by using portal venous phase CT images. Our study demonstrated that a DLS trained by using a large amount of CT data allowed for highly accurate staging of liver fibrosis. In this study, the overall staging accuracy of the DLS was 79.4%. The accuracy of the DLS for diagnosis of significant fibrosis, advanced fibrosis, and cirrhosis was even higher, reaching AUROC values in the range of 0.95–0.97.

The DLS developed in this study was robust across variable clinical settings and imaging conditions; our results indicated that the accuracy of the DLS in staging fibrosis was not dependent on CT scan technique, patient demographic characteristics, or the presence of a liver mass. The only significant independent factor that affected the performance of the DLS was the pathologic fibrosis stage; the diagnosis of intermediate stage fibrosis (ie, F1 and F2) with the DLS was less accurate than the diagnosis of cirrhosis. This result was expected because the changes in liver texture and morphologic structure become more obvious as liver fibrosis progresses (12,16,18). Therefore, CT images of a liver with the intermediate stages of fibrosis may contain imaging features that are less discriminating than those of cirrhosis; however, this observation could reflect the relatively small amount of data for stages F1 and F2 in our development data set even though we tried to avoid such problems by augmentation of image data for these minority classes.

Table 2: Diagnostic Performance of the Deep Learning System for Diagnosing Liver Fibrosis in the Test Data Sets

Diagnostic Performance	Significant Fibrosis	Advanced Fibrosis	Cirrhosis
AUROC	0.96 [0.95, 0.97]	0.97 [0.96, 0.98]	0.95 [0.94, 0.96]
Sensitivity (%)	95.5 (634/664) [93.6, 96.9]	94.6 (476/503) [92.3, 96.4]	84.6 (279/330) [80.2, 88.3]
Specificity (%)	89.9 (204/227) [85.2, 93.5]	95.4 (370/388) [92.8, 97.2]	96.6 (542/561) [94.8, 98.0]
Accuracy (%)	94.1 (838/891) [92.3, 95.5]	95.0 (846/891) [93.3, 96.3]	92.1 (821/891) [90.2, 93.8]

Note.—Data in parentheses are numerator/denominator; data in brackets are 95% confidence interval. AUROC = area under the receiver operating characteristic curve.

Table 3: Influence of Patient Characteristics, CT Data, and Pathologic Data on the Performance of the Deep Learning System for Staging Liver Fibrosis in the Test Data Sets

		Univariate Analysis		Multivariable Analysis†	
Characteristic	DLS Accuracy (%)*	Odds Ratio	*P* Value	Adjusted Odds Ratio	*P* Value
Age					
18–50 y	78.3 (300/383)	1	.51		
>50 y	80.1 (407/508)	1.12 (0.8, 1.55)			
Sex					
Women	79.0 (300/380)	1	.80		
Men	79.7 (407/511)	1.04 (0.75, 1.45)			
Pathologic fibrosis stage					
F0	77.1 (91/118)	0.72 (0.45, 1.16)	.18	0.73 (0.40, 1.35)	.32
F1	71.6 (78/109)	0.55 (0.35, 0.89)	.01	0.51 (0.29, 0.88)	.02
F2	75.2 (121/161)	0.46 (0.28, 0.77)	.003	0.54 (0.33, 0.88)	.01
F3	79.8 (138/173)	0.62 (0.37, 1.04)	.07	0.71 (0.44, 1.14)	.15
F4	84.6 (279/330)	1		1	
Etiologic cause					
HBV	81.0 (323/399)	1		1	
HCV	83.3 (85/102)	1.18 (0.66, 2.10)	.58	1.29 (0.72, 2.33)	.39
Alcohol	79.1 (34/43)	0.89 (0.41, 1.93)	.77	0.90 (0.41, 1.97)	.79
Autoimmune	80.3 (122/152)	0.96 (0.60, 1.53)	.85	1.18 (0.72, 1.94)	.51
Other	73.3 (143/195)	0.65 (0.43, 0.97)	.03	0.81 (0.50, 1.33)	.41
Hepatic tumor					
Absent	78.9 (399/506)	1			
Present	80 (308/385)	1.07 (0.77, 1.49)	.67		
Pathologic specimen					
Biopsy	78.6 (400/509)	1	.52		
Surgery	80.4 (307/382)	1.12 (0.80, 1.55)			
Pathologic data source†					
Internal	80.5 (579/719)	1	.07	1	.26
External	74.4 (128/172)	0.70 (0.48, 1.04)		0.79 (0.52, 1.19)	
CT data source†					
Internal	80.8 (340/421)	1	.32		
External	78.1 (367/470)	0.85 (0.61, 1.18)			
No. of CT multi–detector rows					
1–6	83.3 (120/144)	1	.19		
≥16	78.5 (585/745)	0.73 (0.46, 1.17)			
CT tube voltage					
80–110 kVp	74.3 (52/70)	1	.28		
120–140 kVp	79.7 (653/819)	1.36 (0.78, 2.39)			
CT section thickness					
2–4 mm	76.5 (156/204)	1	.26		
5–10 mm	77.9 (549/685)	1.24 (0.85, 1.81)			

Note.—Unless otherwise indicated, data in parentheses are 95% confidence intervals. DLS = deep learning system.

* Data in parentheses are numerator/denominator.

† The sources of pathologic and CT data were divided into internal (from the same institution as the development data sets) or external groups.

Table 4: Performance of the Deep Learning System, Radiologists, and Serum Fibrosis Tests in Diagnosing Liver Fibrosis in the Test Data Set 1

Parameter	AUROC			
	Significant Fibrosis (n = 303)	Advanced Fibrosis (n = 195)	Cirrhosis (n = 109)	Obuchowski Index
Deep learning system	0.96 (0.94, 0.98)	0.97 (0.95, 0.99)	0.95 (0.93, 0.97)	0.94 (0.93, 0.96)
Radiologist 1	0.78 (0.74, 0.82)	0.83 (0.79, 0.86)	0.86 (0.83, 0.90)	0.77 (0.74, 0.80)
Radiologist 2	0.79 (0.75, 0.83)	0.84 (0.80, 0.88)	0.87 (0.83, 0.90)	0.78 (0.76, 0.81)
Radiologist 3	0.83 (0.79, 0.86)	0.87 (0.83, 0.90)	0.88 (0.84, 0.91)	0.81 (0.78, 0.83)
Radiologist 4	0.75 (0.70, 0.79)	0.80 (0.76, 0.84)	0.84 (0.81, 0.88)	0.74 (0.72, 0.77)
Serum fibrosis test				
APRI	0.85 (0.81, 0.88)	0.72 (0.67, 0.76)	0.65 (0.60, 0.70)	0.71 (0.68, 0.74)
FIB-4	0.84 (0.80, 0.87)	0.78 (0.74, 0.82)	0.76 (0.72, 0.80)	0.76 (0.74, 0.79)

Note.—Data in parentheses are 95% confidence intervals. Data set 1 consisted of 421 patients. Radiologists 1–3 were academic abdominal radiologists; radiologist 4 was a trainee abdominal radiology fellow. APRI = aminotransferase-to-platelet ratio index, AUROC = area under the receiver-operating characteristic curve, FIB-4 = fibrosis-4 index.

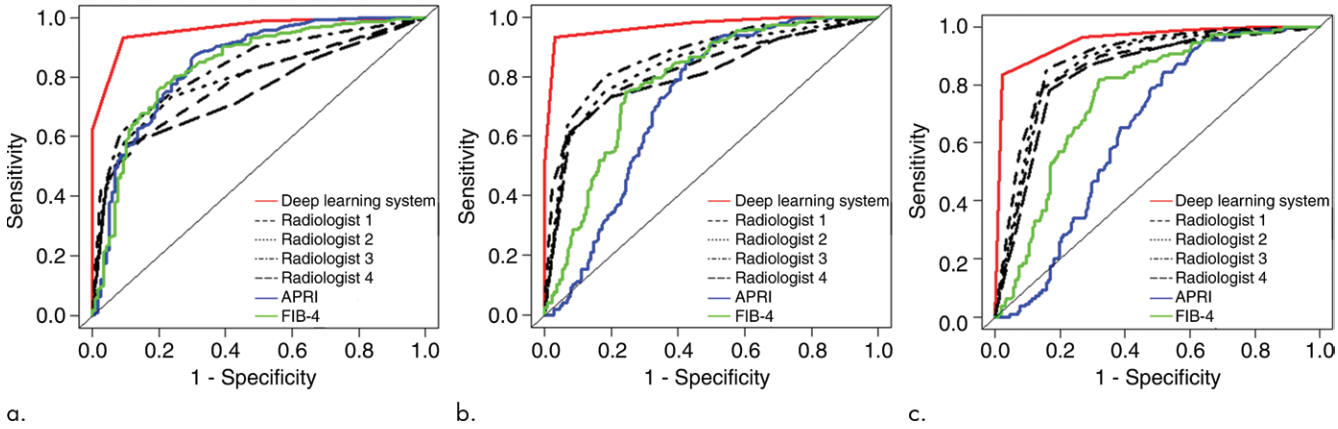

Figure 4: Receiver operating characteristic curves for diagnosis of **(a)** significant fibrosis, **(b)** advanced fibrosis, and **(c)** cirrhosis in test data set 1 with data from 421 patients. The deep learning system achieved areas under the receiver operating characteristic curve of 0.95–0.97 for diagnosis of significant fibrosis, advanced fibrosis, and cirrhosis, and outperformed the four radiologists and the serum fibrosis markers, including aminotransferase-to-platelet ratio index (APRI) and fibrosis-4 index (FIB-4).

In our study, the DLS outperformed radiologists and serum fibrosis tests in the diagnosis of significant fibrosis, advanced fibrosis, and cirrhosis. In clinical practice, staging of liver fibrosis is not routine among radiologists who interpret liver CT images, although the presence of chronic liver disease or cirrhosis is sometimes suggested on the basis of typical imaging findings. This is because CT findings associated with liver fibrosis are subtle in the earlier stage of liver fibrosis, visual assessment of these findings is subjective, and the findings may not occur concordantly. The decision-making process used by a DLS is not traceable, so it is not clearly understood how the DLS could assess liver fibrosis by using CT images better than visual image analyses performed by radiologists. However, we assume that the ability of the DLS to extract and comprehensively analyze numerous features from images may have led to its high accuracy in staging liver fibrosis.

Yasaka et al (24) recently reported on a convolutional neural network system for staging liver fibrosis by using gadoxetic acid–enhanced MR images. The accuracy achieved by our DLS was higher than the AUROC values reported by Yasaka et al of 0.85 or less for diagnosing F2 or greater, F3 or greater, and F4. Furthermore, we developed and validated the DLS by using larger data sets than those in the study by Yasaka et al, who used 534 and 100 data sets for development and validation, respectively. Unlike their system, which used small cropped images of the liver and required radiologists' efforts to capture the liver images, our system processed raw CT images in a fully automated manner.

Among the noninvasive imaging methods available for assessing liver fibrosis, MR- or US-based elastography techniques have been the most extensively validated for their clinical efficacy. Compared with the elastographic techniques, the DLS has the advantage of wide applicability to routine portal venous phase CT images without the need for dedicated scanners or add-on devices. The accuracy of the DLS in assessment of liver fibrosis demonstrated in our study is similar or higher than that of the elastographic techniques (AUROC range, 0.80–0.97 for diagnosing F2–F4, F3–F4, or F4) reported in recent meta-analyses (6,7,9). However, a head-to-head comparison is needed to confirm this finding. Considering the promising results of our study and the widespread use of liver CT in clinical practice, the DLS

a. b. c.

Figure 5: Contrast-enhanced axial CT images of the liver. **(a)** A 53-year-old woman with autoimmune hepatitis and pathologic fibrosis stage of F2. The deep learning system (DLS) correctly predicted the fibrosis stage as F2. The radiologists' visual grades of liver fibrosis were grade 1 for one radiologist, grade 2 for a different radiologist, and grade 3 for two radiologists. The aminotransferase-to-platelet ratio index (APRI) was 2.05 and the fibrosis-4 index was 3.81. **(b)** A 52-year-old woman with chronic B viral hepatitis and a pathologic fibrosis stage of F3. The DLS correctly predicted the fibrosis stage as F3. The radiologists' visual grades of liver fibrosis were grade 1 for two radiologists, grade 2 for one radiologist, and grade 3 for another radiologist. The APRI was 1.7 and fibrosis-4 index was 2.82. **(c)** A 42-year-old man with nonalcoholic steatohepatitis and a pathologic fibrosis stage of F4. The DLS correctly predicted the fibrosis stage as F4. The radiologists' visual grades of liver fibrosis were grade 3 for two radiologists and grade 4 for two radiologists. The APRI was 0.61 and fibrosis-4 index was 1.84.

may have a role as a noninvasive method for assessment of liver fibrosis in patients with chronic liver disease.

Our study had limitations. Our development data set was not balanced for pathologic fibrosis stage and included data from patients with liver tumors. Despite these limitations, our DLS showed high accuracy in staging liver fibrosis. However, the performance of the DLS may have been better if we had trained the DLS with an ideal development data set including a large amount of CT data that was balanced across the different fibrosis stages and obtained from patients with conditions that are clinically relevant in terms of assessment for liver fibrosis. Although we performed the clinical validation of the DLS by using relatively large data sets, the generalizability of this assessment tool needs to be evaluated further. More than half of our test data sets comprised patients with hepatitis B viral or hepatitis C viral, so the DLS should be further validated in patients with nonviral liver diseases, such as nonalcoholic fatty liver disease. Although our test data set included CT data obtained by using various scanners and CT techniques, there has been continual progress in CT techniques, including new image reconstruction algorithms for reduction of the radiation dose (32). Therefore, the applicability of the DLS when novel CT techniques are used should also be evaluated in future studies. Finally, further research would be required to evaluate the clinical benefits of the DLS in predicting the prognosis and helping to guide management of patients with chronic liver disease.

In conclusion, the DLS allows for highly accurate assessment of liver fibrosis by using portal venous phase CT images of the liver. Because of the widespread availability of CT and liver CT imaging, the DLS is a promising and widely applicable method for assessment of liver fibrosis.

Acknowledgments: The authors thank Hyung Cheol Kim, MD, Department of Radiology, Severance Hospital, Yonsei University College of Medicine, for assisting with collection of the imaging data used in this study.

Author contributions: Guarantors of integrity of entire study, S.S.L., H.S.K.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, K.J.C., J.K.J., S.S.L., W.H.S., J.Y.C.; clinical studies, K.J.C., J.K.J., S.S.L., H.S.K., J.Y., J.Y.C., Y.L., B.K.K., J.H.K., S.Y.K., E.S.Y.; experimental studies, Y.S.S., W.H.S.; statistical analysis, K.J.C., J.K.J., S.S.L., W.H.S., H.S.K.; and manuscript editing, K.J.C., J.K.J., S.S.L., H.S.K., J.Y., Y.L., B.K.K., J.H.K., S.Y.K., E.S.Y.

Disclosures of Conflicts of Interest: **K.J.C.** disclosed no relevant relationships. **J.K.J.** disclosed no relevant relationships. **S.S.L.** disclosed no relevant relationships. **Y.S.S.** disclosed no relevant relationships. **W.H.S.** disclosed no relevant relationships. **H.S.K.** disclosed no relevant relationships. **J.Y.** disclosed no relevant relationships. **J.Y.C.** disclosed no relevant relationships. **Y.L.** disclosed no relevant relationships. **B.K.K.** disclosed no relevant relationships. **J.H.K.** disclosed no relevant relationships. **S.Y.K.** disclosed no relevant relationships. **E.S.Y.** disclosed no relevant relationships.

References

1. Standish RA, Cholongitas E, Dhillon A, Burroughs AK, Dhillon AP. An appraisal of the histopathological assessment of liver fibrosis. Gut 2006;55(4):569–578.
2. Bravo AA, Sheth SG, Chopra S. Liver biopsy. N Engl J Med 2001;344(7):495–500.
3. Castera L. Noninvasive methods to assess liver disease in patients with hepatitis B or C. Gastroenterology 2012;142(6):1293–1302.e4.
4. Lubner MG, Malecki K, Kloke J, Ganeshan B, Pickhardt PJ. Texture analysis of the liver at MDCT for assessing hepatic fibrosis. Abdom Radiol (NY) 2017;42(8):2069–2078.
5. Vergniol J, Foucher J, Terrebonne E, et al. Noninvasive tests for fibrosis and liver stiffness predict 5-year outcomes of patients with chronic hepatitis C. Gastroenterology 2011;140(7):1970–1979, 1979.e1–1979.e3.
6. Herrmann E, de Lédinghen V, Cassinotto C, et al. Assessment of biopsy-proven liver fibrosis by two-dimensional shear wave elastography: An individual patient data-based meta-analysis. Hepatology 2018;67(1):260–272.
7. Singh S, Venkatesh SK, Wang Z, et al. Diagnostic performance of magnetic resonance elastography in staging liver fibrosis: a systematic review and meta-analysis of individual participant data. Clin Gastroenterol Hepatol 2015;13(3):440–451.e6.
8. Palmeri ML, Wang MH, Rouze NC, et al. Noninvasive evaluation of hepatic fibrosis using acoustic radiation force-based shear stiffness in patients with nonalcoholic fatty liver disease. J Hepatol 2011;55(3):666–672.
9. Xiao G, Zhu S, Xiao X, Yan L, Yang J, Wu G. Comparison of laboratory tests, ultrasound, or magnetic resonance elastography to detect fibrosis in patients with nonalcoholic fatty liver disease: A meta-analysis. Hepatology 2017;66(5):1486–1501.

10. Pickhardt PJ, Malecki K, Kloke J, Lubner MG. Accuracy of Liver Surface Nodularity Quantification on MDCT as a Noninvasive Biomarker for Staging Hepatic Fibrosis. AJR Am J Roentgenol 2016;207(6):1194–1199.
11. Kim H, Park SH, Kim EK, et al. Histogram analysis of gadoxetic acid-enhanced MRI for quantitative hepatic fibrosis measurement. PLoS One 2014;9(12):e114224.
12. Smith AD, Branch CR, Zand K, et al. Liver Surface Nodularity Quantification from Routine CT Images as a Biomarker for Detection and Evaluation of Cirrhosis. Radiology 2016;280(3):771–781.
13. Besa C, Wagner M, Lo G, et al. Detection of liver fibrosis using qualitative and quantitative MR elastography compared to liver surface nodularity measurement, gadoxetic acid uptake, and serum markers. J Magn Reson Imaging 2018;47(6):1552–1561.
14. Huber A, Ebner L, Heverhagen JT, Christe A. State-of-the-art imaging of liver fibrosis and cirrhosis: A comprehensive review of current applications and future perspectives. Eur J Radiol Open 2015;2:90–100.
15. Venkatesh SK, Yin M, Takahashi N, Glockner JF, Talwalkar JA, Ehman RL. Noninvasive detection of liver fibrosis: MR imaging features vs. MR elastography. Abdom Imaging 2015;40(4):766–775.
16. Bonekamp S, Kamel I, Solga S, Clark J. Can imaging modalities diagnose and stage hepatic fibrosis and cirrhosis accurately? J Hepatol 2009;50(1):17–35.
17. Aguirre DA, Behling CA, Alpert E, Hassanein TI, Sirlin CB. Liver fibrosis: noninvasive diagnosis with double contrast material-enhanced MR imaging. Radiology 2006;239(2):425–437.
18. Lo GC, Besa C, King MJ, et al. Feasibility and reproducibility of liver surface nodularity quantification for the assessment of liver cirrhosis using CT and MRI. Eur J Radiol Open 2017;4:95–100.
19. Goshima S, Kanematsu M, Kobayashi T, et al. Staging hepatic fibrosis: computer-aided analysis of hepatic contours on gadolinium ethoxybenzyl diethylenetriamine-pentaacetic acid-enhanced hepatocyte-phase magnetic resonance imaging. Hepatology 2012;55(1):328–329.
20. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. N Engl J Med 2017;376(26):2507–2509.
21. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA 2017;318(6):517–518.
22. Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA 2016;316(22):2402–2410.
23. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284(2):574–582.
24. Yasaka K, Akai H, Kunimatsu A, Abe O, Kiryu S. Liver Fibrosis: Deep Convolutional Neural Network for Staging by Using Gadoxetic Acid-enhanced Hepatobiliary Phase MR Images. Radiology 2018;287(1):146–155.
25. Yu E; Korean Study Group for the Pathology of Digestive Diseases. Histologic grading and staging of chronic hepatitis: on the basis of standardized guideline proposed by the Korean Study Group for the Pathology of Digestive Diseases [in Korean]. Taehan Kan Hakhoe Chi 2003;9(1):42–46.
26. Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology 1996;24(2):289–293.
27. Wai CT, Greenson JK, Fontana RJ, et al. A simple noninvasive index can predict both significant fibrosis and cirrhosis in patients with chronic hepatitis C. Hepatology 2003;38(2):518–526.
28. Sterling RK, Lissen E, Clumeck N, et al. Development of a simple noninvasive index to predict significant fibrosis in patients with HIV/HCV coinfection. Hepatology 2006;43(6):1317–1325.
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44(3):837–845.
30. Lambert J, Halfon P, Penaranda G, Bedossa P, Cacoub P, Carrat F. How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. Clin Chem 2008;54(8):1372–1378.
31. Nguyen P. NonbinROC: software for evaluating diagnostic accuracies with non-binary gold standards. J Stat Softw 2007;21(10):1–10.
32. Shuman WP, Chan KT, Busey JM, et al. Standard and reduced radiation dose liver CT images: adaptive statistical iterative reconstruction versus model-based iterative reconstruction-comparison of findings and image quality. Radiology 2014;273(3):793–800.